Pund-IT

PUND-IT RESEARCH

Weekly Review

August 2, 2006

In this issue:

- **Moving to the Head of the Data Class**
  By David G. Hill, The Mesabi Group

Pund-IT, Inc.
2776 Sulphur Drive
Hayward, CA
U.S.A. 94541

Phone: 510-909-0750
Fax: 510-886-4937
charles@pund-it.com
www.pund-it.com

Moving to the Head of the Data Class

By David G. Hill, Mesabi Group

Information lifecycle management (ILM) is the policy-driven process of managing information as it changes value throughout the full range of its lifecycle from conception to disposition. For ILM to become a reality, however, businesses must classify the data they wish to manage. Identifying and ordering data according to business and regulatory requirements requires tools that use a policy-management engine based upon essential business rules, as well as metadata and content knowledge of files and/or databases. As a result, organizations can classify data on the basis of value or requirements such as compliance or availability.

The benefits of data classification reflect and bolster the benefits of ILM: more efficient use of the storage infrastructure through the use of tiered storage solutions, greater productivity for storage management, enabling or simplifying compliance management and eDiscovery processes, enabling information that was lost to be found and used effectively. Developing greater knowledge of information enables querying across broader sets of data, identification of new relationships between data for competitive advantage, easier programming at a higher (business metadata) level, better compliance (legal discovery can find the information needed), enhanced data quality (if you use ETL tools), and better data administration across data stores.

Getting Users and Management Committed Is the First Challenge

A number of software tools have sprung up to help automate the process of data classification. These tools are typically policy-management-driven, which means that there is a policy "engine" that applies business rules. This sounds great, but the software machinery can only mesh the gears of automation after business users have engaged business rules. And therein lies the rub.

Business rules are set by the organization, specifically IT and executive management with strategic responsibility for business information, but getting these groups and individuals involved is not always easy. First, they need a compelling reason to participate, which means that they have to benefit directly from the process. The fact that data classification organizes data so that IT can manage it better should warm the hearts of IT management, but does little for business managers. For example, although tiering of storage allows IT (and therefore the enterprise as a whole) to save money, business management is unlikely to see any compelling direct benefit (even if there is a sophisticated chargeback scheme where the cost savings appear).

Compliance *might* be one "killer" application for data classification, since it is a business task that can attract the buy-in of executives responsible for legal discovery and compliance audits. However, compliance tends to focus on one or just a few applications so it is not a compelling motivation for universal data classification. Still it is a good start. A second "killer" application *might* be enterprise search, since a good deal of information that has value to the enterprise is often misplaced or forgotten. Search is also an important component for eDiscovery as well.

Even if IT can get management to participate, setting business process rules is not easy even for the willing. For example, trying to determine the meaning of simple terms such as "customer" or "product" across cross functional boundaries can be a major challenge. However, there are good semi-automated tools that can help to create data classification metadata where there was none before.

Deciding on the Right Data Classification Tools

A couple of the terms that have popped up to cover the data classification space are "information classification and management" and "intelligent information management." Both are good attempts at product categorization, but some vendors decline to be pigeon-holed. Categorization is an attempt to view products through a single lens not only for comparison purposes, but also to help IT

organizations understand what they need to get their hands around data classification and what solutions they can buy for that process.

The products listed (Table 1) are from companies that focus on the data classification space in some general sense. Storage resource management (SRM) tools or classification tools that are only a part of a larger package are not listed.

Table 1: Sampler of Data Classification Vendors

Vendor	Product	Product Focus	Technology Foundation
Abrevity	FileData Classifier FileData Manager	Adds file content visibility and file tagging to file system-created metadata to provide what it calls "information value management" to manage data including true unstructured data	Claims that its SliceBASE data model is more efficient than relational database and enterprise search tools with a scalability of more than one billion files
Arkivio	Arkivio autostor	Focuses on information lifecycle management functions including data discovery, classification, and migration	ILM-policy management for retention, compliance, archiving, and consolidation that, among other functions, uses algorithms that monitor the change of data value and storage value over time
Index Engines	Index Engines Search and Classification Appliance	A search engine that spans all internal islands of indexing to enable what it calls "dynamic classification."	Appliance that provides a full, storage-efficient indexing service on top of the storage infrastructure
Kazeon	Kazeon Information Server — IS1200 family	Content-aware classification and policy-based management of files distributed across an enterprise	LAN-based, agent-less, out-of-band file management appliance for full content indexing and search across all distributed managed files
Njini Software	Njini IAM Suite	Indexes, categorizes, and single instances data. Focuses on value of information and uses the term information asset management (IAM).	Active real-time inline policy engine through which all data must pass (in-band) in order to collect the necessary information and create metadata for managing the data
Scentric	Scentric Destiny	Universal classification across all data types from structured to unstructured	Built around a distributed enterprise-wide catalog that stores key data attributes and content indexes for all data. Data can be classified into logical groups and managed via policies.
StoredIQ	StoredIQ's Information Classification and Management Platform	Content-based discovery and classification of information for a business or security policy for targeted vertical industries	Appliance that extracts keywords, patterns, and natural language concepts and entities rather than creating a full text index

Source: Mesabi Group, August 2006

The data classification market is still very fluid and products/partnerships continue to be announced. The companies listed in the table represent smaller vendors, but most have partnerships with some of the larger players in the industry. Abrevity lists among its technology partnerships one with the EMC Velocity Partner program and a Gold Business Partner relationship with HP. Among its industry partners Arkivio includes EMC, Hitachi Data Systems, NetApp, and Symantec. Index Engines states that EMC and NetApp are among its technology partners. Kazeon declares that

Hitachi Data Systems and NetApp are but two of its partners. Scentric lists Hitachi Data Systems as a technical partner. StoredIQ touts EMC and NetApp as technology partners. Note that the partnerships are by no means exclusive. Note also the absence of some key IT vendors, such as IBM and Sun Microsystems, in these lists.

Looking at Data Classification Products through Different Lenses

Enterprises can look through a couple of perceptual lenses to help them determine which data classification solution may serve their needs. Vendor products tend to be a composite of functions and the function sets are not the same for each product.

The Management Lens

The first filter is to determine which types of management functions are performed: storage, data, or information. The three types (derived from the Storage Networking Industry Association or SNIA) are:

Storage management — discovers, monitors, and controls physical storage assets.

Data management — the non-data-path control and use of the data itself from creation to deletion, such as migration, replication, and backup/restore processes.

Information management — manages the content and decision-making relationships of information as it moves through the lifecycle of a business process, such as records management and content management.

What is the difference? Storage management covers tiering, data management focuses on data protection (such as employing different types of data protection for different classes of data) and migration, and information management is about content-awareness (what is inside is what is important), such as applying eDiscovery.

In a broader sense, information management is the enterprise-wide administration at the metadata/business level across all vendors/data types. There can also be a mix of types working in harmony and integration. File metadata (a data management function) may be mixed with an index of information (a content-aware information management function) to classify data. That classified data can then be migrated (a data management function) to the appropriate tier of storage (a storage management function).

Storage management is at the block level and uses primitive metadata, such as when the block was last accessed. Data management can use file and database metadata (it is at the level of the file or the database "record," but does not understand the content of the file or record). Information management is content-aware in that the contents of a file or database can be examined and that information (either directly or in the form of an index) can be used for classification purposes.

Products that fall within the "information classification and management" or "intelligent information management" categorization scheme may have functions that fall within the different categories. For example, one product may focus on migration of data to tiers of storage as well as classification. Another product may focus on classification using both file metadata and content-based search capability, but not do migration.

The Data Lens

The second way of looking at data classification is through the types of data that the data classification process manages. Data classification does not have to be universal for an enterprise. One application at a time or a series of interrelated applications can be selected. However, data classification may involve only one data type or a mix of data types.

Table 2: Differentiating Among the Different Types of Data

Type	Structured	Semi-structured	Unstructured
Common Forms	Database	"Text" documents, such as e-mail, word processing, presentations, spreadsheets	Natively bitmapped data, such as video, audio, pictures, and MRI scans
Key Differentiator	Sort	Search	Sense
Examples	OLTP systems, such as CRM and ERP Data warehousing	Personal productivity, such as e-mail and word processing Web sites using HTTP	Entertainment, such as video and audio Imaging, such as digital photography and bitmapped medical tests

Source: Mesabi Group, August 2006

The most common differentiation is between structured and unstructured data. What users typically consider to be structured data — data in databases — is essentially correct. What is frequently considered unstructured data — for example, where word processing documents are commingled with video files — is not a correct categorization. There is an essential differentiation between semi-structured and unstructured data in that semi-structured data can be effectively searched. For example, for example, one can search for all e-mails or word processing documents (i.e. those supported by content-aware applications) that contain a certain word. That is why there is a need for the semi-structured category.

The same search capability cannot be applied to native unstructured data (Note the use of the word native.) For example, questioning when was a certain word spoken in a movie is unanswerable in native mode since video cannot be searched but only sensed (viewed or heard). Speech recognition can be used to determine whether and when a word was spoken and this information can then be put into a searchable format. The goal for a lot of unstructured data is to increase its structure by pairing it with complementary structured or semi-structured information.

The term semi-structured is most often used to refer to e-mail, while other semi-structured data is relegated to unstructured status. While e-mail is semi-structured so are word processing documents. What probably separated these classes of documents is that the mistaken impressions of vendors thinking of Microsoft Exchange, which is a composite application that contains multiple data types. However, to achieve optimum data classification success the focus should be on the data. There is nothing intrinsic in an e-mail that gives it more "structure" than a word processing document.

Since the word unstructured tends to be used gratuitously (and inaccurately), a determination has to be made between unstructured and semi-structured data. The distinction is important. True unstructured data cannot be used natively by content-aware applications. Unstructured data is typically stored in BLOBs (Binary Large Objects), which, of course, are changed less. Thus administration and classification have to be different.

Note that no one product has to encompass all types of data; only that all types of data within the organization have to be covered if universal data classification is an appropriate goal for the enterprise. This is a classic Caveat Emptor example for IT customers, who must look and consider carefully the types of data supported by the software tools under consideration.

Mission Accomplished?
Many useful products are available to help businesses with the data classification process. At the upcoming Storage Networking World conference, SNIA will be holding a Hands On Lab to enable end users to be able to "test drive" a variety of different data classification products. Helping potential customers develop a feel for available tools should help them gets their arms (and heads) around a difficult and complex process. Organizations have to be able to determine the types of management functions they need to perform (such as storage management for tiering, data management for migration, and content-based search for information management). Enterprises also need to decide what data (structured, semi-structured, and unstructured) needs to be classified as the tools may be proficient at one data type or multiple data types.

Still the challenge remains on how to get commitment from business users and management to engage in the data classification process. One way of achieving that goal may be to leverage and integrate the concepts of ILM with enterprise content management, master data management, and business process management. That way the business will be able to understand the benefits of data classification in ways that are important for the efficient and effective operation of business functions and business units. Then, data classification will lead to true intelligent information management, i.e. blending the content and decision-making relationships of information throughout the lifecycle of the business process.

About The Mesabi Group
The Mesabi Group helps organizations make their complex storage, storage management, and interrelated IT infrastructure decisions easier by making the choices simpler and clearer to understand.